变量的相关性分析与建模

参考资料：

[1] 《R语言医学数据分析实践》

[2] 《人工智能（AI）在微生物组中的应用-随机森林分类与回归预测》（[链接](https://mp.weixin.qq.com/s?__biz=MzIyNzIyNTczNA==&mid=2247491475&idx=1&sn=8b287625fc3b81f4a3e7b62f48c5e4d8&chksm=e865267adf12af6c8635c6515d5210a097c02a28301dc8923504593611d69db3d80ae2814a93&mpshare=1&scene=1&srcid=0111yoLaQNqztt7uDBpDxMBn&sharer_sharetime=1665219182692&sharer_shareid=a21bc68461a00a088b357021b5f54298&exportkey=n_ChQIAhIQznvpRTYQGqSeoISoWeicMRKZAgIE97dBBAEAAAAAABm7EKECpWAAAAAOpnltbLcz9gKNyK89dVj0lVx7hVdJUQzSruoUII4PN4Fgau6UnuWI88uCZNFqaPhyRjUbOa5JaagXcy6L89z9Jh2DaZkSJjC6pjp6eXjT18JNhFiJiqlNu%2F74Gs8tzIgKOul2R4lFVNtnI7ESdbzbF3Nnc78xS4VvAd2vcbMFb15jvtj27BpbllZ3umtJljzdmTSu1uXOfOLFKPGHnVwasXFg1SnhgypoxZ59y9KsOmcGbS7T60rsaBPTxuVpM6AwYJ6ejZK07TpfqLIN40jJSvBVgiNH7UCWpXCS%2Bdi0Ent%2Fbdm0PqAFirkz3EQG2PJNPiiKD1HwHbrt9cpMeF1qErOc&acctmode=0&pass_ticket=YJ66UFlQihrcwAB%2BWKLAUbh5d1jkuRdJTLrhXWG2sda6OQtg%2Bwbji4yE9uvkR3%2FX&wx_header=0#rd)）

一 两变量之间的相关系数 （[1] 111-115页）

（1）cor()函数

cor(dataframe, method=”pearson”)

注：

（a）dataframe每一行一个样本，每一列一个变量

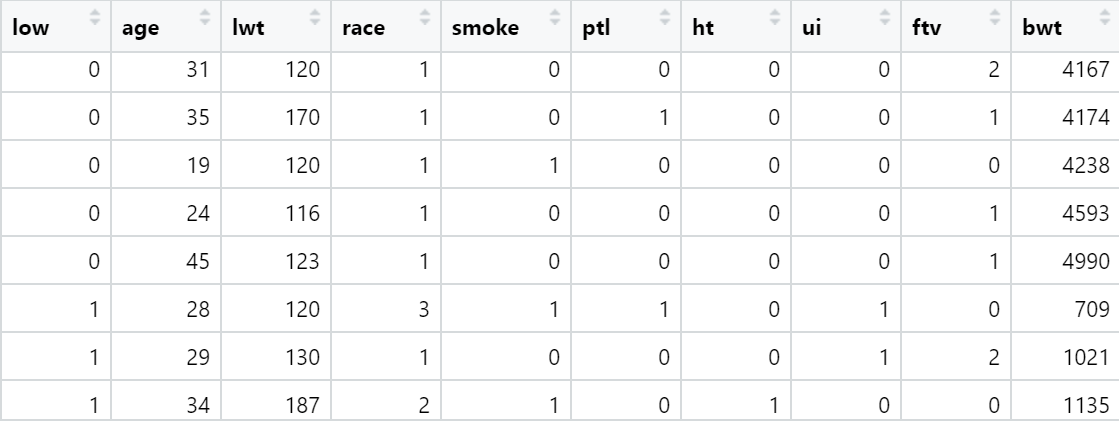
（b）method可选项：“pearson”, “spearman”, “kendall”, 注意每个方法适用范围

（2）psych包corr.test()函数

corr.test(dataframe, method=”pearson”)

分析例子：birthwt数据集

是否低体重 母亲年龄 母亲体重 种族 是否吸烟 早产次数 高血压 子宫应激 孕前就医次数 新生儿体重



（3）gmm包的pcor()函数和pcor.test()函数进行偏相关分析

（4）卡方检验中分类变量的相关性：vcd包中的assocstats()函数和epiDisplay包中的kap()函数

二 数据建模

输出

拟合算法

特征输入

两组概念：

（1）分类 vs 回归 （2）训练 vs 测试

三 线性回归 ([1]第6章)

Y=α+β1X1+β2X2+…+βnXn+ε

（1）拟合函数（训练）：lm()

（2）结果展示：summary()

（3）变量重要性评估：drop1()

（4）变量选择（特征选择）：step()

（5）预测（测试）：predict.lm()

（6）回归诊断：gvlma()

（7）汇总输出（写文章要展示的结果）：regress.display()

分析例子：

（1）大骨节病患儿的年龄与其尿肌酐含量的研究。

（2）ISwR包中的cystfibr数据集，囊泡性纤维癌患者肺功能研究。

女性1男性0 体质指数 呼气量



四 逻辑回归 ([1]第7章)

log(P/(1-P))=α+β1X1+β2X2+…+βnXn

（拟合的是优势比的对数）

（1）拟合函数（训练）：glm()

（2）结果展示：summary()

（3）变量重要性评估：drop1()

（4）变量选择（特征选择）：step()

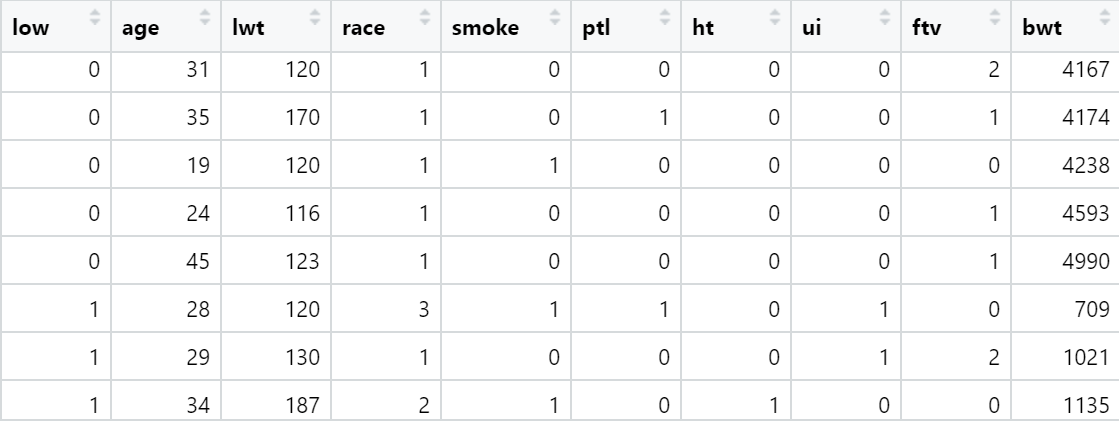
（5）预测（测试）：predict()

（6）汇总输出（写文章要展示的结果）：logistic.display()，（epiDisplay）包

注：上述函数主要是二分类，多分类略有不同

二分类分析例子：birthwt数据集

是否低体重 母亲年龄 母亲体重 种族 是否吸烟 早产次数 高血压 子宫应激 孕前就医次数 新生儿体重



因子转化：

（1）早产次数超过1次的孕妇很少，把ptl转化成二分类因子，无早产与有早产

（2）探访医生超过1次的孕妇很少，把ftv转化成三分类因子，即无探访、探访1次、探访2次以上

glm1 <- glm(low~age+lwt+race+smoke+ptl+ht+ui+ftv,data=birthwt,family = binomial)

summary(glm1)

注：对多分类变量的处理方式：

方法一：把n分类变量变成n个二分类变量

white: [0,0,1]; black: [0,1,0]; other: [1,0,0]

方法二：把方法一简化一下

white: [0,0]; black: [0,1]; other: [1,0]

log(P/(1-P))= 0.82302-0.03723×age-0.01565×lwt+1.19241×black+0.74069×other+0.75553×smoke+1.34376×ptl+1.91317×ht+0.68019×ui-0.43638×ftv1+ 0.17901×ftv2

多分类分析例子：流产史、怀孕次数对孕妇生理状态的影响

孕妇生理状态（三分类） 流产史 怀孕次数



五 随机森林 ([2])

（1）训练：randomForest包的randomForest()函数

（2）观察变量的重要性：$ importance

（3）预测：predict()